Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs

Year: 2023 | Citations: 351 | Authors: Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han

---

Abstract

In this study, we introduce adaptive KV cache compression, a plug-and-play method that reduces the memory footprint of generative inference for Large Language Models (LLMs). Different from the conventional KV cache that retains key and value vectors for all context tokens, we conduct targeted profiling to discern the intrinsic structure of attention modules. Based on the recognized structure, we then construct the KV cache in an adaptive manner: evicting long-range contexts on attention heads emphasizing local contexts, discarding non-special tokens on attention heads centered on special tokens, and only employing the standard KV cache for attention heads that broadly attend to all tokens. Moreover, with the lightweight attention profiling used to guide the construction of the adaptive KV cache, FastGen can be deployed without resource-intensive fine-tuning or re-training. In our experiments across various asks, FastGen demonstrates substantial reduction on GPU memory consumption with negligible generation quality loss. We will release our code and the compatible CUDA kernel for reproducibility.